Neural Reasoning – A quick and dirty tutorial introduction

Learning to Reason with Language and Logic
Using Statistical and Symbolic Techniques

Vijay Saraswat

IBM Research
August 2017

Why do we need to revisit Logic?

IBM

Modern Mathematical Logic (~ 1879 Begriffschrift +)

- Propositions, Crisp individuation, Quantification – expressing recognized structure-in-the-world via formal symbols
- Models, Truth
- Inference, Proof
- Proof-search, Proofs-as-Types
- Constraints – satisficing, optimizing in crisp domains
- Ontologies, Formal Mathematics

Modern Machine Learning (~ 2008 +)

- Techniques for recognizing structure-in-the-world from data, generalizing beyond given (labeled) data, becoming better over time (through interaction)
- Techniques to leverage massive amounts of (noisy, uncertain) data
- Vector (distributed) representations for text, capturing a novel ("associational") approach to meaning

Learning to Reason with Language and Logic
Integrating Statistical and Symbolic techniques

- Unified representation for language and logic – concepts expressible symbolically and through (learnt) distributed representations
- Enabling reasoning at different levels of precision / ambiguity – text, formulas
- Based on a principled (probabilistic) approach to uncertainty
- Supporting a variety of professional level tasks in deep information domains

Example professional-level tasks

Scenario Planning

Determine consistency and plausibility of a scenario, given its textual description, and causal model of the world.

Business strategy

Continuous Awareness

Analyze news for information that affects probability distribution across projected scenarios.

Business strategy, investment planning, health

Diagnosis

Work with a customer calling in to determine nature of his/her service problem and its remediation.

Call center support

Obligation Extraction

Determine, from newly received regulatory text, changes in obligations for the given enterprise, and consequences for existing business processes.

Compliance

Constraint validation

Determine whether a proposed course of business action is valid, given known compliance constraints.

Compliance

Quote Generation and Understanding

For web users, generate insurance policy quotes from given data, and answer questions about it.

Compliance

Example task: Constraint Violation

Article 4: 1. Without prejudice to Directive 2000/53/EC, Member States shall prohibit the placing on the market of:
(a)all batteries or accumulators, whether or not incorporated into appliances, that contain more than 0,0005% of mercury by weight; …
2. The prohibition set out in paragraph 1(a) shall not apply to button cells with a mercury content of no more than 2 % by weight.

DIRECTIVE 2006/66/EC

R: '_ prohibits the placing on the market of _'(State, Item):-
R=rule('Directive 2006/66/EC', ['Article 4', 1, a]),
'member state'(eu,State),
'battery or accumulator'(Item),
applicable(R,Item),
'mercury content'(Item,'by weight',X percent),
{X > 0.0005}.

Without prejudice to Directive 2000/53/EC, Member States shall prohibit the placing on the market of all batteries or accumulators, whether or not incorporated into appliances, that contain more than 0,0005% of mercury by weight;

Member States shall prohibit the placing on the market of all batteries or accumulators, that contain more than 0,0005% of mercury by weight.

Tyranny of language – many different ways of saying the same thing

Utterance may have irrelevant information

Relevant information may be spread across sentence-chunks: resolve anaphora.

Vijay Saraswat 12:13 AM
@watson Can I use battery M6512 in a new product in Europe? It has 0.16% mercury, and also contains Cadmium and other rare metals.

Respond at appropriate level of generality

Yes, if it is a button cell. (Or, perhaps, in some other cases.) Do you want details?

Implicit context:
"in the future"
"determine all applicable regulations"

Reason with numerical relationships

Actually it is more complex

DIRECTIVE 2013/56/EU

Article 1
Directive 2006/66/EC is amended as follows:
(1) Article 4 is amended as follows:
(a) paragraph 2 is replaced by the following:
'2. The prohibition set out in paragraph 1(a) shall not apply to button cells with a mercury content of no more than 2 % by weight until 1 October 2015.';
…

Subsequent regulations invalidate some portions of old regulation.

Regulation does this by simply replacing clauses in old regulations with new clauses.

Statistical techniques do not understand "quotes"!
Logical techniques can. Vital to use both.

Technical Challenge: Construct the currently active obligations compositionally*

* Compositionality: Conjunction of obligations extracted separately from Reg A and Reg B should give the right result even if Reg B rupdates Reg A.

Long-term Vision -- Upton: Achieving Professional Facility in Deep Domains

IBM

> "[Edison] was brimming over with ideas but needed someone with advanced mathematical skills who could do calculations and research the scientific literature to help solve intractable problems. Despite his inveterate suspicion of academic scientists, Edison found Upton highly engaging and quite useful."

Professional facility: Operate as an assistant with the mastery of professionals in the field

Deep domains: Domains with significant amount of pre-existing (formalizable) knowledge

"..Francis Robbins Upton, the very first student to officially earn, by examination, a graduate degree from Princeton. He received a Masters of Science in 1877."

A bot capable of earning a PhD from a major university.

Deep and challenging AI goals – without committing to "General AI"!

Upton: Key principles

- Support continuous ingestion of documents
  - System always has latest data – attractive to users.
  - Support joint learning of text and visual content
- Early on, get a usable (cloud-connected) system in the hands of real users, designed for continuous feedback.
  - Users are legal, compliance, financial professionals – not linguists / computer scientists or KR engineers.
  - Build system around key functionality critical for users
  - Support **teams** – with appropriate work-flow
  - Ensure users can correct / provide feedback for every system response, in terms that make sense to them.
  - Do this early to maximize learnings – putting a-man-in-the-box initially, if necessary.

- Design system **to learn continuously**

- **Support** automated conversion from input text to multiple structured forms, including linguistic and logical forms
  - Learn to reason with language **and** logic

7

Upton: Task-based (Statistical + Symbolic) Reasoning Architecture

Interaction/User Experience / Workflow

Traditional UX and Workflow

Knowledge Extraction
Dialog management

Machine learning

Natural Language Generator

Natural lang generation
(template-driven,
or process driven)
Visualization generation

Machine learning

Quasi-logical forms

Task input

Knowledge Store & DB

Background theories, ontologies, constraints, rules, transformations, Lexicons, sentence patterns
Textual Theories

Data-bases

Task output

Problem-solvers/Reasoners

Constraint-solvers
Logical Reasoners
(abductive, probabilistic)

Deep-learning classifiers, predictors, mappers, reasoners

Explanation generators
Optimization Engines
Analytic Engines
Code generators
Graph Reasoners
Planners
Simulators

Machine learning

Task Descriptions

Auxiliary sources (e.g. answer store, quote generator)

Domain-specific Enrichment
Extraction, Resolution & Integration of Entity, Relationship and Numeric Data
Continuous and Active learning, Mixed-initiative human curation.

Machine learning

Knowledge Extractor

Continuous Ingestion of documents

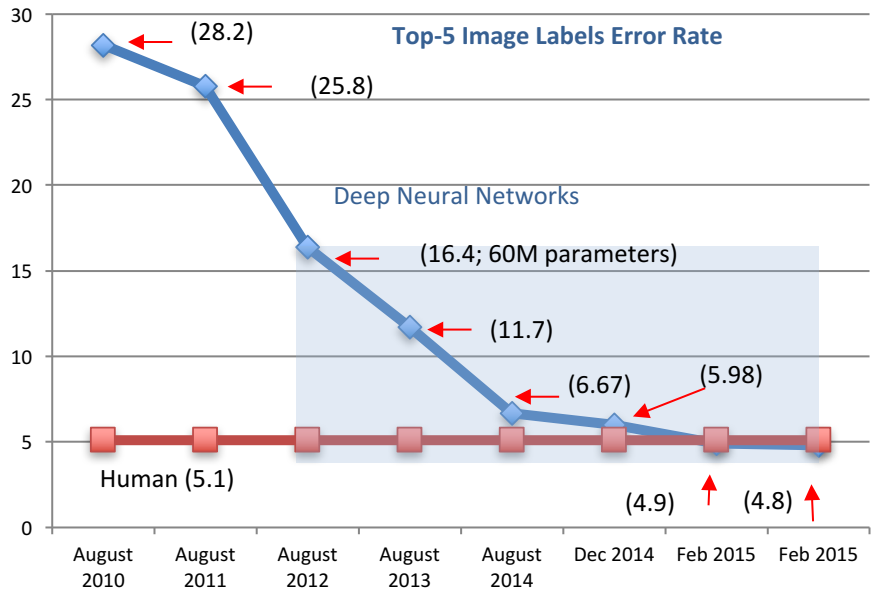
Content Acquirer

Modern Machine Learning

A very quick introduction

Machine Learning: Breakthrough performance in Image Recognition

Over 40 NN deployed in production.
Google query ranking (page-rank, multiple algorithms, regression)
Video recommenders for Youtube
Photosearch for Google+
Android speech recognition
Google News – ranking, clustering
Robot navigation, path planning



Spam detection
Face recognition
Ordering of contacts
Ordering of stories in newsfeed
Ad presentation
Estimating popularity of news features
Clustering posts

Dialog – Siri
Recommender systems (Amazon, Pandora)
Machine Translation (Skype)

Product classification (Walmart)
Classification (starts, galaxies, exo-planets, …)
Categorize products in posted images

Prediction of hard disk failure (Baidu)
Trending, who to follow (Twitter)

Anomaly detection (US insurance companies, Paypal)

Nvidia Tesla revenue from DL: 0% -- 2013, 25% 2014, 33% (est) 2015

Machine Learning is here, now.

Supervised Machine Learning, via gradient descent, for NN

Space of all models $\mathcal{F}$

$\mathcal{L}(f_T(\underline{x}),\underline{y})$ Loss function

1. Initialize T appropriately
2. For each observation $(\underline{x},y)_k$, compute loss
3. Use partial derivatives of the loss function wrt each parameter t in T to update parameter value in direction of descent
4. Iterate until convergence.

f_T

Parameterized model

$(\underline{x},\underline{y})_k$

Two pass algorithm:

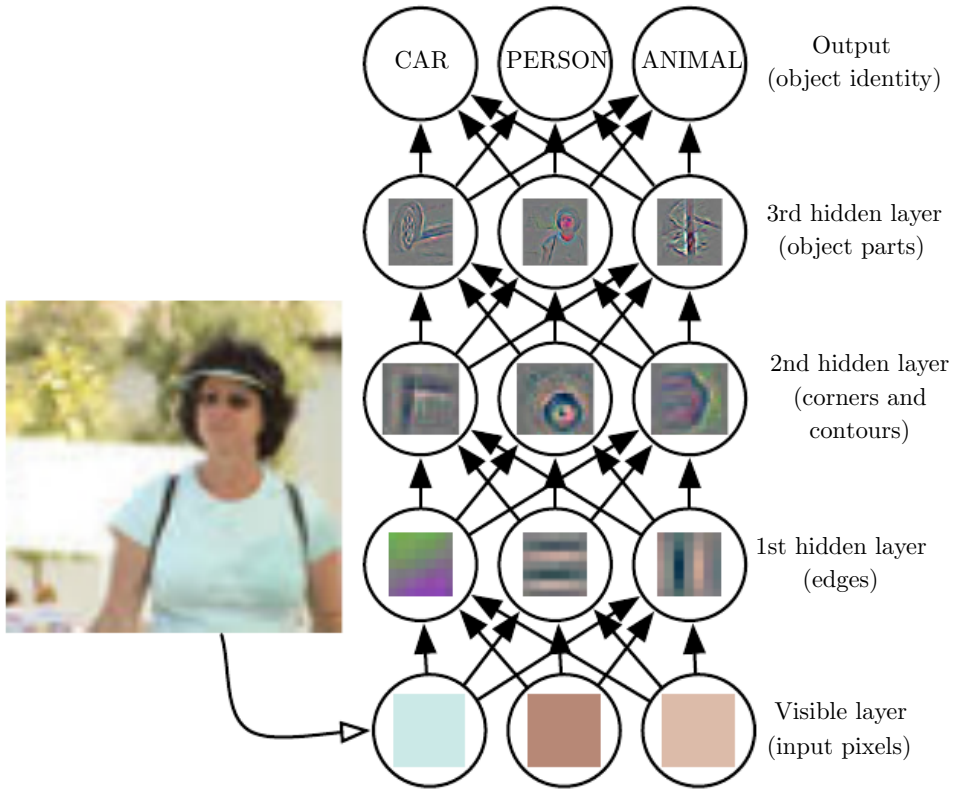
(a) Forward: compute $f_T(x)$, given T and x.

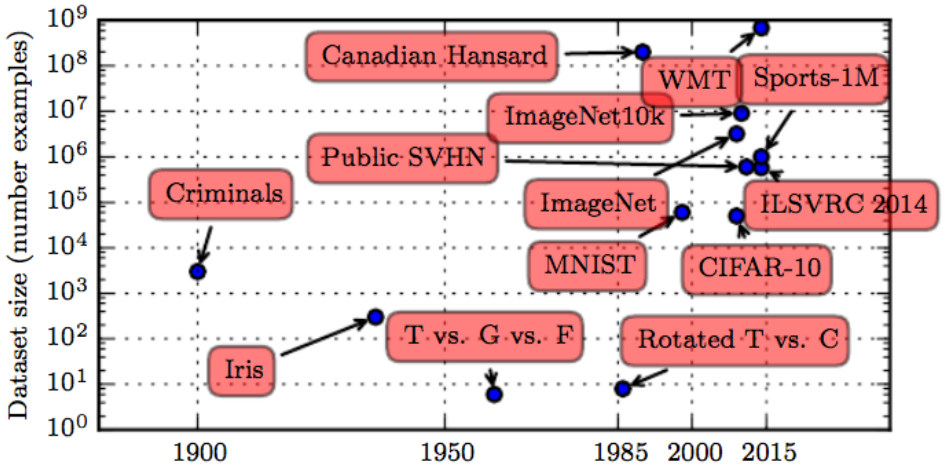
(b) Backward: compute gradients for each parameter in T. Apply.

GPU-friendly ~ $O(P^3)$

A practical approach for non-convex optimization

DNNs – A quick introduction

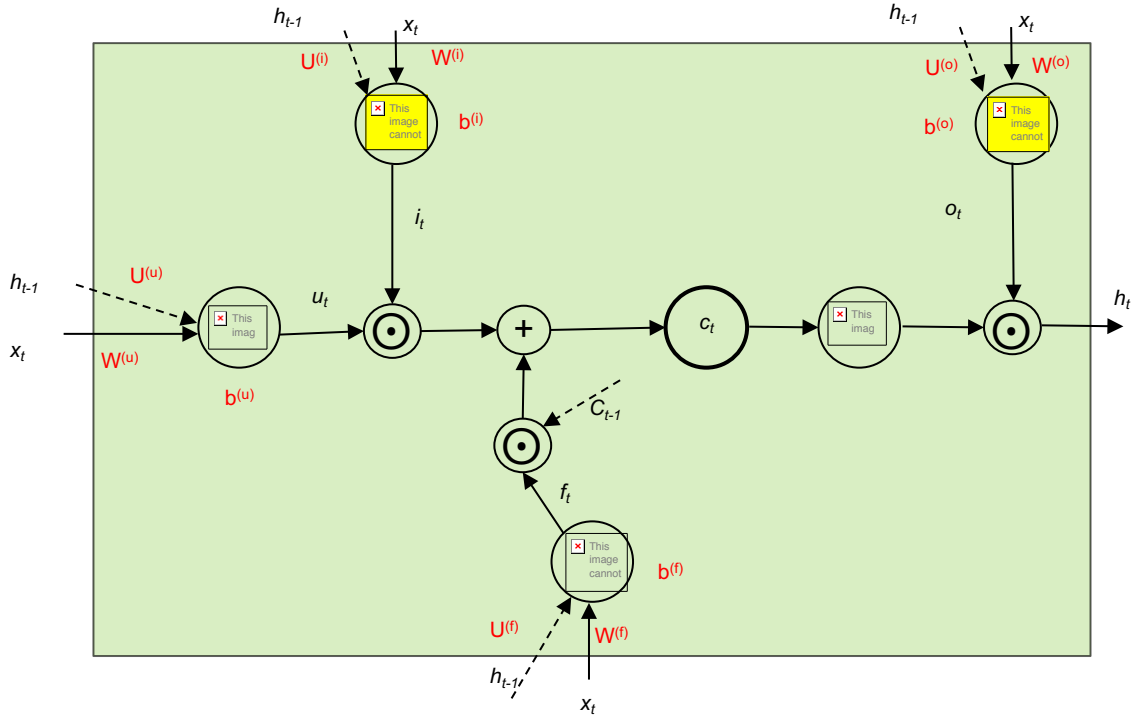




- Multi-layer (>= 100 in some cases), very high parameter (~ billions) neural networks (computing elements: linear and some non-linear functions)
- Usually leverage distributed representations of symbols
- Trained end-to-end with gradient descent techniques → little need for manual feature-construction
- Of interest in supervised, unsupervised and reinforcement learning settings

Goodfellow, Bengio, Courville. 2016. *Deep Learning*
Patel, Nguyen, Baraniuk, 2015. *A Probabilistic Theory of Deep Learning.*

LSTMN: Dealing with recurrence, using explicit gating on input, output, memory



input at current time-step

Hidden state generated at current time-step

logistic sigmoid fn (=$1/(1+e^{-x})$)

$$i_t = \sigma\left(W^{(i)}x_t + U^{(i)}h_{t-1} + b^{(i)}\right), \quad (1)$$

Input gate $i_t \in [0,1]^d$

$$f_t = \sigma\left(W^{(f)}x_t + U^{(f)}h_{t-1} + b^{(f)}\right),$$

Forget gate $f_t \in [0,1]^d$

$$o_t = \sigma\left(W^{(o)}x_t + U^{(o)}h_{t-1} + b^{(o)}\right),$$

Output gate $o_t \in [0,1]^d$

$$u_t = \tanh\left(W^{(u)}x_t + U^{(u)}h_{t-1} + b^{(u)}\right),$$

$$c_t = i_t \odot u_t + f_t \odot c_{t-1},$$

Memory cell c_t

$$h_t = o_t \odot \tanh(c_t),$$

Hidden state h_t

Elementwise multiplication

All vectors in $\Re^d$

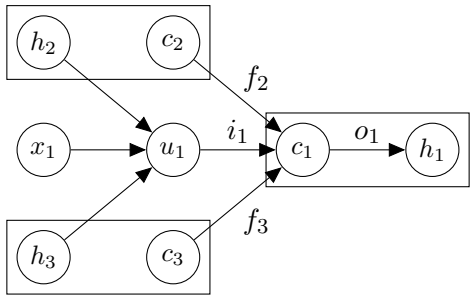
$W^?$, $U^?$, $b^?$: trained parameters

- RNNs introduce ability to deal with recurrence but are hard to train.
- LSTMNs introduce *forget*, *input* and *output* gates:
  - Forget controls the extent to which the previous memory cell is forgotten
  - Input controls how much each unit is updated
  - Output gate controls the exposure of the internal memory state

Zaremba, Sutskever. 2014. *Learning to execute*

Hochreiter, Schmidhuber. 1997. *Long Short-Term Memory. Neural Computation* 9(8):1735–1780.

Dealing with branching tree structure: N-ary Tree-LSTMNs



(input, output, forgetting) gate and memory updates now depend on states of many children units
+ one forget gate per child

$$i_j = \sigma\left(W^{(i)}x_j + \sum_{\ell=1}^{N} U_\ell^{(i)} h_{j\ell} + b^{(i)}\right), \quad (9)$$

$$f_{jk} = \sigma\left(W^{(f)}x_j + \sum_{\ell=1}^{N} U_{k\ell}^{(f)} h_{j\ell} + b^{(f)}\right), \quad (10)$$

(k in 1..n)

$$o_j = \sigma\left(W^{(o)}x_j + \sum_{\ell=1}^{N} U_\ell^{(o)} h_{j\ell} + b^{(o)}\right), \quad (11)$$

$$u_j = \tanh\left(W^{(u)}x_j + \sum_{\ell=1}^{N} U_\ell^{(u)} h_{j\ell} + b^{(u)}\right), \quad (12)$$

$$c_j = i_j \odot u_j + \sum_{\ell=1}^{N} f_{j\ell} \odot c_{j\ell}, \quad (13)$$

$$h_j = o_j \odot \tanh(c_j), \quad (14)$$

Reduces to LSTMNs for sequences

Note: training is complicated because the structure of the net depends on the structure of the tree, which depends on the training example (cf DyNet from CMU).

Working with tree-structured inputs critical for reasoning applications (axioms are trees)

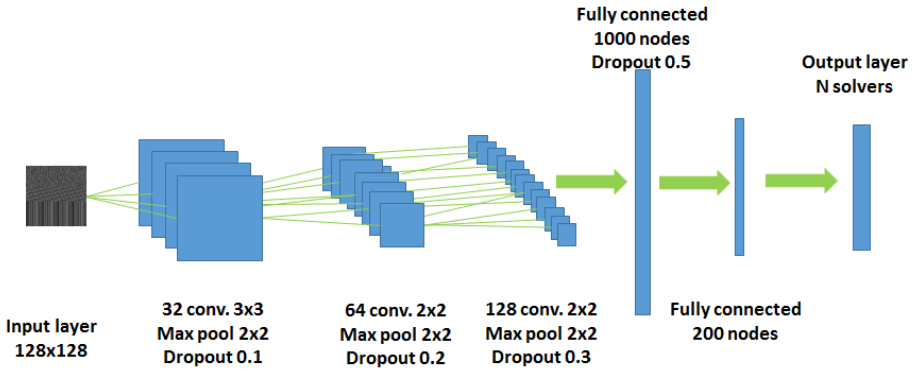
Tai, Socher, Manning. 2015. *Improved Semantic Representation from Tree-Structured LSTMNs*

Applications in Reasoning

Can performance of Reasoners be improved using Deep Learning techniques?

Yes. Replace manual feature engineering for key heuristic steps with deep learning.

Algorithm Selection without Feature Engineering



Dataset	CSHC	BSS	CNN	New Feat.
Industrial	95.6 ± 2.56	92.6 ± 2.50	93.0 ± 2.73	93.7 ± 2.56
Random	96.5 ± 1.89	77.4 ± 1.74	90.7 ± 1.90	90.1 ± 1.45
Crafted	91.6 ± 2.45	77.2 ± 4.75	78.2 ± 6.32	79.9 ± 4.11
CSP	94.5 ± 2.94	65.4 ± 4.98	82.1 ± 2.31	78.2 ± 2.97

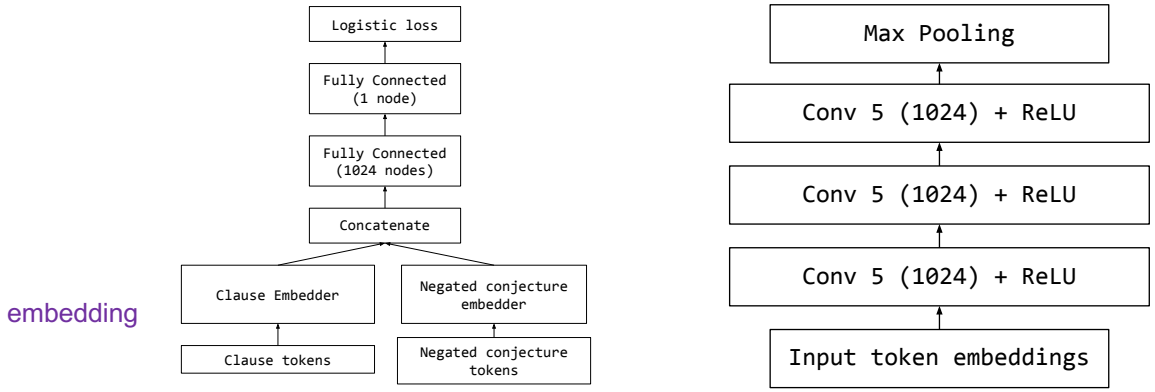
This paper

- Problem: Selecting which solver to use for a particular problem instance (CSP, SAT) is determined heuristically, based on manually generated problem features.
  - Idea: Train a DNN to select best solver for given instance.
- Problem: Problem instances are of variable size
  - Solution: View problem as an image, that can be scaled up and down, apply standard convolution NN techniques.
  - *Today: Likely use a Tree LSTMN to encode the problem; should improve results.*

Loreggia, Malitsky, Samulowitz, Saraswat, 2015. *Deep Learning for Algorithm Portfolios*

Automated Theorem Proving: DNNs for Premise Selection

Also used WaveNet and Tree-LSTM



Train on existing ATP proofs (32, 524 / 57,917)

Model	DeepMath 1	DeepMath 2	Union of 1 and 2
Auto	578	581	674
*WaveNet 640	644	612	767
*WaveNet 256	692	712	864
WaveNet 640	629	685	997
*CNN	905	812	1,057
CNN	839	935	1,101
Total (unique)	1,451	1,458	1,712

New proofs found for 1,712/25,361 of hard MML theorems (did not have ATP generated proofs before).
(20 GPUs, using Keras and Tenserflow, with Adam optimizer)

- MaLeCoP [Urban2011] showed using ML to choose next step can lead to 20x reduction in steps. Here: (State-of-the-art, saturation-based, FOL theorem prover) E, operating on Mizhar Mathematical Library (large repository of theorems, >= 50K)
- Problem: Improve proof search performance using a DNN by better predicting whether a given clause is needed for proof of a given proposition. (No feature engineering!)
- Solution:
  - Use two phase approach: DNN-guided phase followed by fast heuristics.
  - Embed clauses into dense vectors, leverage definitions.
- Paper evaluates multiple DNN architectures: convolutional, WaveNet, Tree-LSTMNs

Loos, Irving, Szegedy, Kaliszyk, 2017. *Deep Network Guided Proof Search*
Alemi, Chollet, Een, Irving, Szegedy, Urban, 2017. *DeepMath – Deep Sequence Models for Premise Selection*
Kaliszyk, Chollet, Szegedy. HolStep, 2017: *A Machine Learning Dataset for Higher-order Logic Theorem Proving*

Applications in Reasoning

Learning to Reason with Language

Towards a unification of symbolic and differentiable representations

Reasoning in embedding-based KBs

…

A language model assigns a probability to a sequence of words:

$$p(w_1, \ldots, w_m) = \prod_{i=1}^{m} p(w_i \mid w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} p(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1})$$

These are of great value in various tasks such as speech recognition, language translation, information retrieval.

The individual probabilities can be calculated from n-gram counts:

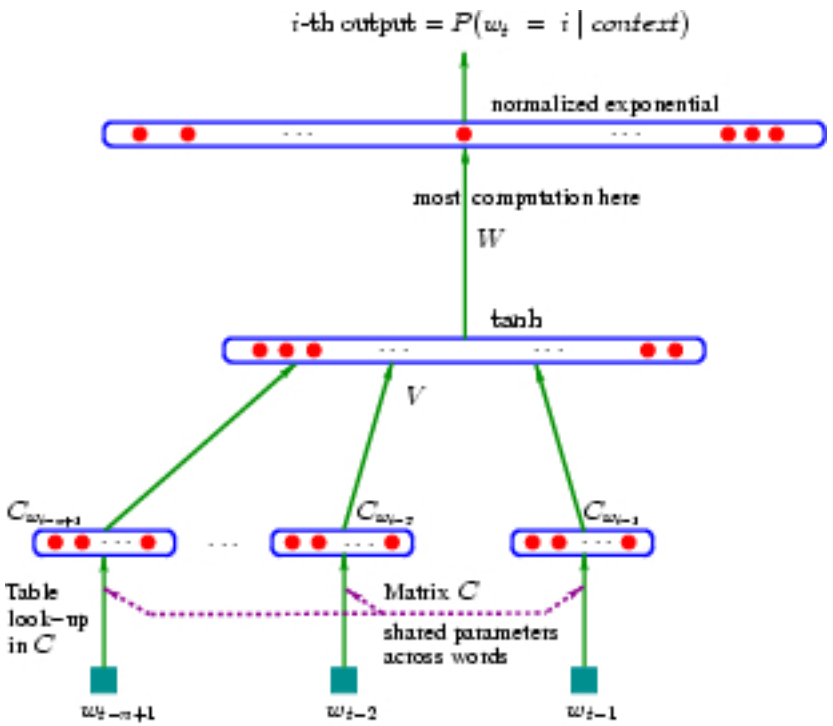
$$\prod_{i=1}^{m} p(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1}) = \frac{count(w_{i-(n-1)}, \ldots, w_i)}{count(w_{i-(n-1)}, \ldots, w_{i-1})}$$

Suffers from curse of dimensionality: not enough data available as n increases. (Increase in n necessary to capture long-range dependencies)

Also does not consider similarity of words, e.g. *dog* and *cat* in the context *A _ was running in a room*.

Language Models / word vecs address a critical problem in applying NN-based ML ideas: how do you internalize symbols as vectors?

Neural Language Models: learn and use distributional representations for words



$softmax(z)_j = e^{z_j} \Big/ \sum_k e^{z_k}$

$p(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1})$ is obtained by embedding each word into a vector in $\Re^d$, then using a standard NN probabilistic classifier on the ($(n-1)d$ dimensional) concatenation x of these vectors.

$$p(w_i = k \mid w_{i-(n-1)}, \ldots, w_{i-1}) = softmax(\boldsymbol{a})_k$$

$$a_k = b_k + \sum_{i=1}^{h} W_{ki} \tanh(c_i + \sum_{j=1}^{(n-1)d} V_{ij}\, x_j)$$

Parameters $\theta = (b, W, c, V)$

Hyper-parameters:

h -- # hidden units

d -- # learned word features

Parameters learnt by stochastic gradient descent to maximize log−likelihood:

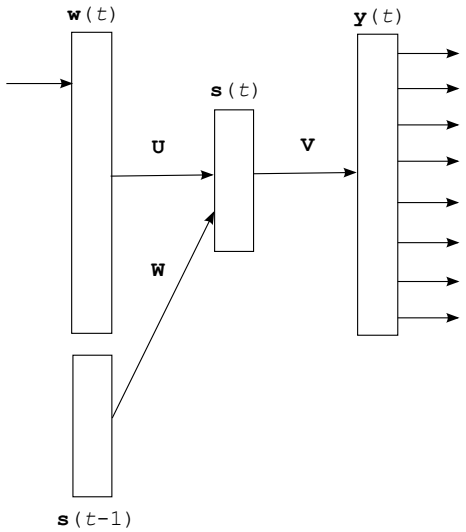
$$L(\theta) = \sum_t \log p(w_t \mid w_{t-n+1}, \ldots, w_{t-1})$$

Dense Vector Representations learn similarities present in data

Bengio et al, 2001, 2003. *A Neural Probabilistic Language Model*.

Mikolov, 2012. *Statistical Language Models based on Neural Networks*

Neural Language Models: with RNNs



sigmoid

$$s_j(t) = f\left(\sum_i w_i(t)u_{ji} + \sum_l s_l(t-1)w_{jl}\right) \quad (3.1)$$

$$y_k(t) = g\left(\sum_j s_j(t)v_{kj}\right) \quad (3.2)$$

softmax

- Feed-forward NN effectively produces an NN version of an N-gram model.
- Introducing an RNN permits entire past history to be digested.
- More complex patterns can also be remembered.

- Several variations developed:
  - Continuous bag-of-words model
  - Skip-gram: predict word in context (at a given position) from given word.

- Large improvements in accuracy on word-similarity, reduced computation costs.

Mikolov, 2012. *Statistical Language Models based on Neural Networks*
Mikolov, Chen, Corrado, Dean. 2013. Efficient estimation of word representations in vector space.
Mikolov, Sutskever, Chen, Corrado, Dean. 2013. *Distributed Representations of Words and Phrases and their Compositionality*
Arora, Li, Liang, Ma, Risteski. 2016. *RAND-WALK: A latent variable model approach to word embeddings*

Textual Entailment

Natural Language ”Reasoning”

Textual Entailment

T: X offers to rent to Y an upstairs floor for a cabaret. Thereafter, the city adopts a fire code making use of the premises illegal without substantial rebuilding.
H: Renting the premises is now illegal.

T D advertises a reward of $100 for the return of his pet dog. G, unaware of the offer, returns D's dog.
H G did not have knowledge of the offer.

T This offer is only available to people 18 years of age and older. Carlos is 17. Carlos heard the offer.
H This offer is not available to Carlos.

T Offeror mails a written offer to offeree stating that acceptance is valid only if received by the offeror within ten days. Offeree mails back the acceptance within ten days but it arrives late.
H The acceptance is late.

More examples

t47
d advertises a reward of $ 100 for the return of his pet dog . g , unaware of the offer , returns d 's dog .
g did not have knowledge of the offer .
entailment

t48 d advertises a reward of $ 100 for the return of his pet dog . g , unaware of the offer , returns d 's dog .
g had knowledge of the offer .
Contradiction

t77
a licensed real estate agent jokingly offers to sell your home for a 1 % commission .
the promise was made in apparent jest .
entailment

t78
a licensed real estate agent jokingly offers to sell your home for a 1 % commission .
the promise was not made in apparent jest .
contradiction

t158
s gives b a stereo today . b promises to pay s $ 500 in one week .
there is promise offered by b and accepted by s as inducement to enter into an agreement .
entailment

Examples from prior corpora: RTE

1-8-H:
Crude oil for April delivery traded at $37.80 a barrel, down 28 cents
Crude oil prices rose to $37.80 per barrel
Neutral

2-12-H:
Oracle had fought to keep the forms from being released
Oracle released a confidential document
neutral

3-13-H:
iTunes software has seen strong sales in Europe.
Strong sales for iTunes in Europe.
entailment

4-15-H:
All genetically modified food, including soya or maize oil produced from GM soya and maize, and food ingredients, must be labelled.
Companies selling genetically modified foods don't need labels.
neutral

5-19-H:
Researchers at the Harvard School of Public Health say that people who drink coffee may be doing a lot more than keeping themselves awake - this kind of consumption apparently also can help reduce the risk of diseases.
Coffee drinking has health benefits.
entailment

Examples from prior corpora: SICK

188
Various people are eating at red tables in a crowded restaurant with purple lights
A small group of people is waiting to eat in a restaurant
neutral

192
A motorcycle rider is standing up on the seat of a white motorcycle
No motorcycle rider is standing up on the seat of a motorcycle
contradiction

194
Nobody is on a motorcycle and is standing on the seat
Someone is on a black and white motorcycle and is standing on the seat
contradiction

200
A motorcyclist is riding a motorbike dangerously along a roadway
A motorcyclist is riding a motorbike along a roadway
entailment

201
There is no motorcyclist riding a motorbike along a roadway
A motorcyclist is riding a motorbike along a roadway
contradiction

291
Several children are sitting down and have their knees raised
Several children are lying down and are raising their knees
neutral

293
Two young girls are sitting on the ground
Two girls are sitting on the ground
entailment

296
Two girls are lying on the ground
Several children are sitting down and have their knees raised
neutral

301
A nude lady is walking in front of a crowd in body paint
There is no lady walking in body paint in front of a crowd
contradiction

302
A nude lady is walking in front of a crowd in body paint
A nude crowd in body paint is walking in front of a lady
neutral

Examples from prior corpora: fracas

024
Many delegates obtained interesting results from the survey.
Many delegates obtained results from the survey.
entailment

025
Several delegates got the results published in major national newspapers.
Several delegates got the results published.
entailment

026
Most Europeans are resident in Europe.
Most Europeans can travel freely within Europe.
entailment

027
A few committee members are from Sweden.
At least a few committee members are from Scandinavia.
entailment

028
Few committee members are from Portugal.
There are few committee members from southern Europe.
neutral

138
Every report has a cover page.
Smith signed the cover page of R-95-103.
entailment

139
A company director awarded himself a large payrise.
A company director has awarded and been awarded a payrise.
entailment

140
John said Bill had hurt himself.
John said Bill had been hurt.
entailment

141
John said Bill had hurt himself.
Someone said John had been hurt.
neutral

142
John spoke to Mary.
Bill spoke to Mary.
Entailment

Examples from prior corpora: SNLI

3691765410.jpg#1r1e
A man with facial hair and a red and gray shirt tugging on a piece of rope.
A man has facial hair.
entailment

6160193920.jpg#1r1c
A middle-aged oriental woman in a green headscarf and blue shirt is flashing a giant smile.
The middle aged oriental woman is watchingt v
contradiction

6160193920.jpg#1r1e
A middle-aged oriental woman in a green headscarf and blue shirt is flashing a giant smile.
A middle aged oriental woman in a green headscarf and blue shirt is flashing a giant smile
entailment

6160193920.jpg#1r1n
A middle-aged oriental woman in a green headscarf and blue shirt is flashing a giant smile.
The middle aged oriental woman is very happy
neutral

6160193920.jpg#0r1n
An Asian woman in a blue top and green headscarf smiling widely as another woman rows a boat in the background.
An Asian woman is happy because she found money on the ground.
neutral

3706019259.jpg#2r1n
A family with a baby, the father is wearing a save the children sign.
A woman is holding a baby.
neutral

3706019259.jpg#3r3c
A foreign family is walking along a dirt path next to the water.
they are riding a bike
contradiction

3706019259.jpg#3r1n
A foreign family is walking along a dirt path next to the water.
People are walking next to a lake.
neutral

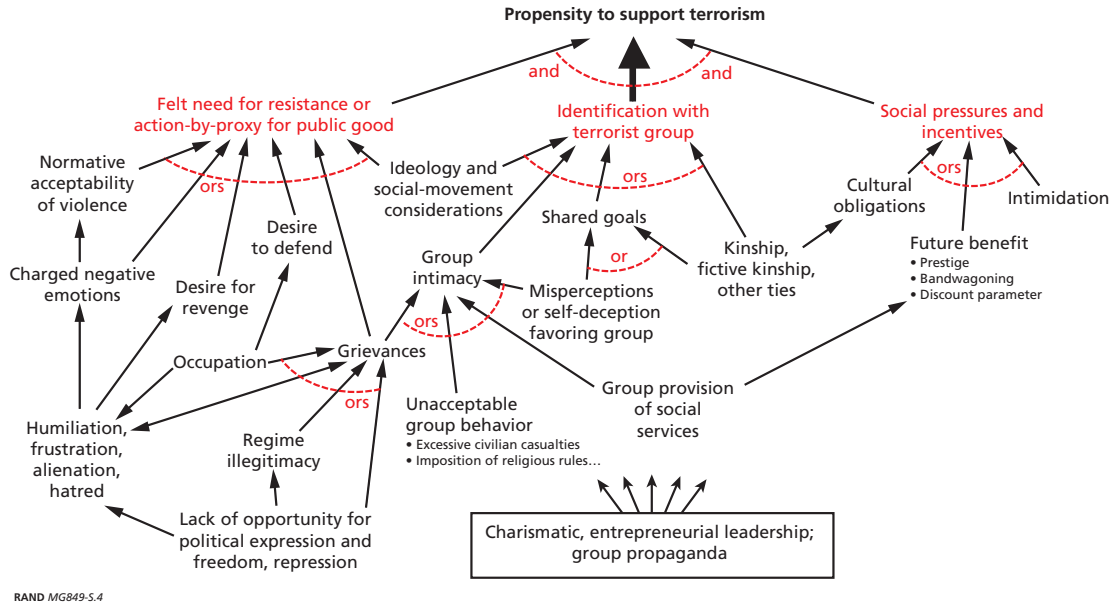
3706019259.jpg#3r4n
A foreign family is walking along a dirt path next to the water.
A foreigner group of cousins hike along a trail next to a stream.
neutral

SNLI Benchmark: Current results

Publication	Model	Parameters	Train (% acc)	Test (% acc)
Feature-based models				
Bowman et al. '15	Unlexicalized features		49.4	50.4
Bowman et al. '15	+ Unigram and bigram features		99.7	*78.2*
Sentence encoding-based models				
Bowman et al. '15	100D LSTM encoders	220k	84.8	77.6
Bowman et al. '16	300D LSTM encoders	3.0m	83.9	80.6
Vendrov et al. '15	1024D GRU encoders w/ unsupervised 'skip-thoughts' pre-training	15m	98.8	81.4
Mou et al. '15	300D Tree-based CNN encoders	3.5m	83.3	82.1
Bowman et al. '16	300D SPINN-PI encoders	3.7m	89.2	83.2
Yang Liu et al. '16	600D (300+300) BiLSTM encoders	2.0m	86.4	83.3
Munkhdalai & Yu '16b	300D NTI-SLSTM-LSTM encoders	4.0m	82.5	83.4
Yang Liu et al. '16	600D (300+300) BiLSTM encoders with intra-attention	2.8m	84.5	*84.2*
Munkhdalai & Yu '16a	300D NSE encoders	3.0m	86.2	*84.6*
Other neural network models				
Rocktäschel et al. '15	100D LSTMs w/ word-by-word attention	250k	85.3	83.5
Pengfei Liu et al. '16a	100D DF-LSTM	320k	85.2	84.6
Yang Liu et al. '16	600D (300+300) BiLSTM encoders with intra-attention and symbolic preproc.	2.8m	85.9	85.0
Pengfei Liu et al. '16b	50D stacked TC-LSTMs	190k	86.7	85.1
Munkhdalai & Yu '16a	300D MMA-NSE encoders with attention	3.2m	86.9	85.4
Wang & Jiang '15	300D mLSTM word-by-word attention model	1.9m	92.0	86.1
Cheng et al. '16	300D LSTMN with deep attention fusion	1.7m	87.3	85.7
Cheng et al. '16	450D LSTMN with deep attention fusion	3.4m	88.5	86.3
Parikh et al. '16	200D decomposable attention model	380k	89.5	86.3
Parikh et al. '16	200D decomposable attention model with intra-sentence attention	580k	90.5	86.8
Munkhdalai & Yu '16b	300D Full tree matching NTI-SLSTM-LSTM w/ global attention	3.2m	88.5	87.3
Wang et al. '17	BiMPM	1.6m	90.9	87.5
Sha et al. '16	300D re-read LSTM	2.0m	90.7	87.5
Chen et al. '16	600D ESIM + 300D Syntactic TreeLSTM (code)	7.7m	93.5	**88.6**
Wang et al. '17	BiMPM **Ensemble**	6.4m	93.2	**88.8**

Extraction of Causal Models

Examples of Existing Causal Models in Social Sciences



Why and How Some People Become Terrorists.
Todd C. Helmus

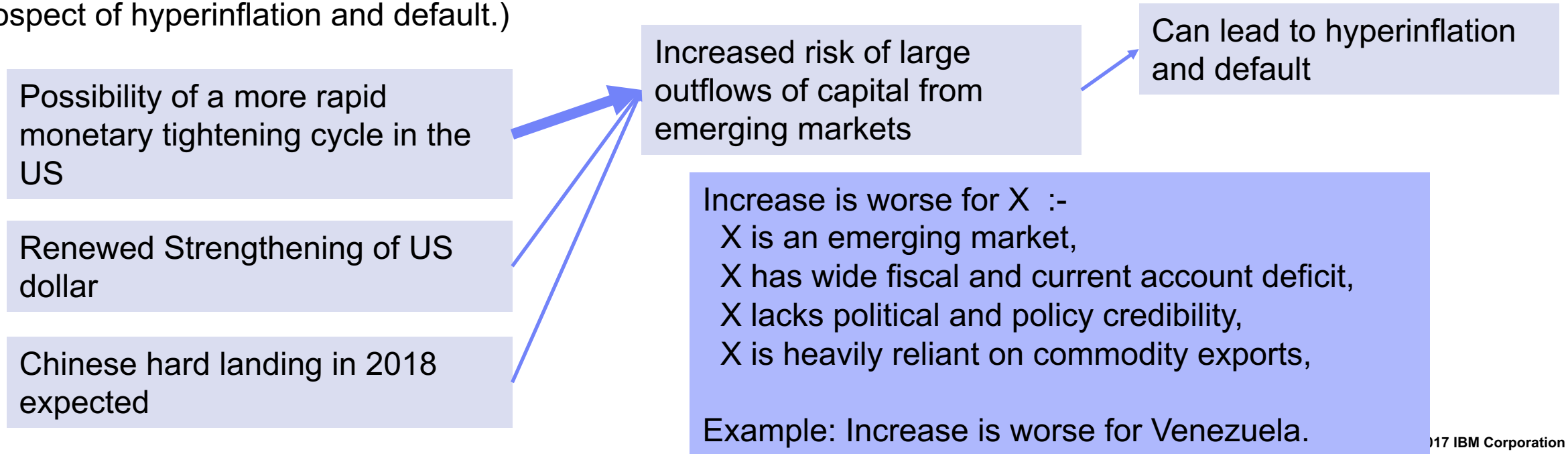
in Rand 2009 study "Social Science for Counterterrorism"

the likelihood that terrorism will ensue as a result of root causes will increase if the social group in question believes that violence is legitimate (even if others see it as terrorism), if it has substantial motivations (perhaps stemming from grievances), *and* if social structures exist permitting the terrorist actions. To a first approximation, however, all three factors are *necessary,* as indicated by the "ands."

the acceptability of terrorism may be driven by a cultural propensity for violence, by ideology (including but not necessarily religion), by political repression and regime illegitimacy, *or* by foreign occupation. e operative word is "or." None of these are *necessary*. Any one might be *sufficient*, or it might be that combinations of two or more of them would be necessary. One factor may substitute for another.

Examples of Existing Causal Models in Social Sciences

The possibility of a more rapid monetary tightening cycle in the US following the election of Mr Trump, coupled with the renewed strengthening of the US dollar and our expectation of a Chinese hard landing in 2018, has increased the risk of large outflows of capital from emerging markets to safer investments. The countries most vulnerable to tighter US monetary policy are those with wide fiscal and current account deficits; those viewed as lacking political and policy credibility; and those heavily reliant on commodity exports. (In the case of Venezuela, all three, combined with policy shortcomings, have raised the prospect of hyperinflation and default.)



Significant existing work in extracting causal models

[Khoo 1998]

[Girju 2003]

Automatic detection of causal relations for QA
Seminal work, used inductive learning (C4.5 decision trees) to generate rules. (.74 precision .89 recall on a small test – identify causal relations of a particular kind in text.

[Chang 2004]

Used unsupervised techniques on large corpus to learn statistics

Survey in [Asghar 2016]

See COPA task [Gordon 2012]

[Karinsky 2012] Pundit

Learning to predict (causal reasoning with text) – uses 150 years of NYTimes headlines + ontologies, to construct generalized rules (used in prediction).

Observation (given)	Prediction
Magnitude 6.5 earthquake rocks the Solomon Islands	*A tsunami warning will be issued for the Pacific Ocean*
Cocaine found at Kennedy Space Center	A few people will be arrested

[Zhao 2016]

Use statistical techniques for connective analysis

[Luo 2016]

Commonsense Causal Reasoning between short texts [Luo et al KR'16]

The COPA Challenge (Choice of Plausible Alternatives)

Example 1 *Premise:* I knocked on my neighbor's door. *What happened as an effect?*
Alternative 1: My neighbor invited me in.
Alternative 2: My neighbor left her house.

Premise: The man broke his toe. What was the CAUSE of this?
Alternative 1: He got a hole in his sock.
Alternative 2: He dropped a hammer on his foot.

Premise: I tipped the bottle. What happened as a RESULT?
Alternative 1: The liquid in the bottle froze.
Alternative 2: The liquid in the bottle poured out.

Researchers' goal was to pursue **coverage** and **precision** – so work with large data, and develop deeper techniques to identify causality in text.

Worked with 1.6B (Bing) web pages (10TB)
Extracted 62,675,002 edges
Over 64,436 nodes (lemmatized terms) (41% of the words in WordNet)

Table 2: Top necessary and sufficient causal pairs

Necessary Causal Pairs	Sufficient Causal Pairs
$(\text{man}_c, \text{kidnapping}_e)$	$(\text{neuroma}_c, \text{pain}_e)$
$(\text{man}_c, \text{jolliness}_e)$	$(\text{eyestrain}_c, \text{headache}_e)$
$(\text{wind}_c, \text{corkscrew}_e)$	$(\text{flashlight}_c, \text{light}_e)$
$(\text{rainfall}_c, \text{flooding}_e)$	$(\text{typhoon}_c, \text{damage}_e)$
$(\text{accident}_c, \text{snarl}_e)$	$(\text{sunrise}_c, \text{light}_e)$
$(\text{erosion}_c, \text{rill}_e)$	$(\text{claustrophobia}_c, \text{panic}_e)$
$(\text{crash}_c, \text{gash}_e)$	$(\text{quake}_c, \text{damage}_e)$
$(\text{virus}_c, \text{tonsillitis}_e)$	$(\text{bacteria}_c, \text{meningitis}_e)$
$(\text{fight}_c, \text{carnage}_e)$	$(\text{quake}_c, \text{loss}_e)$
$(\text{earthquake}_c, \text{avalanche}_e)$	$(\text{overproduction}_c, \text{growth}_e)$

Table 3: COPA results comparison

Data Source	Methods	Accuracy(%)
Web corpus	PMI (W=5)	61.6%
	PMI (W=10)	61.0%
	PMI (W=15)	60.4%
	PMI (W=25)	61.2%
	$CS_{\lambda=0.5}$	**64.8%**
Gutenberg	PMI (W=5)	58.8%
	PMI (W=25)	58.6%
LDC Gigaword	UTDHLT Bigram PMI	61.8%
	UTDHLT SVM	63.4%
ConceptNet	Fuzzy match	51.3%
1-Million Stories	PMI (W=25)	65.2%
10-Million Stories	PMI (W=25)	**65.4%**
CausaNet	$CS_{\lambda=1.0}$	**70.2** %

Commonsense Causal Reasoning between short texts [Luo et al KR'16]

Intended to answer COPA challenge
Example 1 *Premise:* I knocked on my neighbor's door.
What happened as an effect?
Alternative 1: My neighbor invited me in.
Alternative 2: My neighbor left her house.

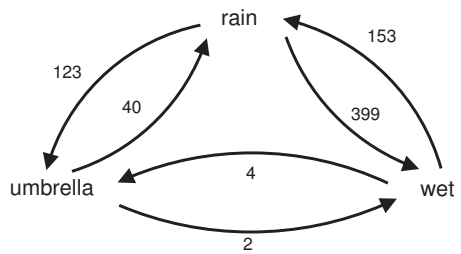


Figure 1: A fragment of causal network

Table 1: 53 Causal cues. *A* is a cause span, and *B* is an effect span. DET stands for a/an/the/one. BE stands for is/are/was/were.

intra-sentence			inter-sentence		
A lead to B	A leads to B	A led to B	If A, then B	If A, B	B, because A
A leading to B	A give rise to B	A gave rise to B	B because A	B because of A	Because A, B
A given rise to B	A giving rise to B	A induce B	A, thus B	A, therefore B	B, A as a consequence
A inducing B	A induces B	A induced B	Inasmuch as A, B	B, inasmuch as A	In consequence of A, B
A cause B	A causing B	A causes B	B due to A	Due to A, B	B in consequence of A
A caused B	B caused by A	A bring on B	B owing to A	B as a result of A	As a consequence of A, B
A brought on B	A bringing on B	A brings on B	A and hence B	Owing to A, B	B as a consequence of A
B result from A	B resulting from A	B results from A	A, hence B	A, consequently B	A and consequently B
B resulted from A	the reason(s) for/of B BE A		A, for this reason alone , B		
DET effect of A BE B	A BE DET reason(s) of/for B				

$$CS(i_c, j_e) = CS_{nec}(i_c, j_e)^{\lambda} CS_{suf}(i_c, j_e)^{1-\lambda} \quad (6)$$

$$CS_T(T_1, T_2) = \frac{1}{|T_1| + |T_2|} \sum_{i \in T_1} \sum_{j \in T_2} CS(i, j) \quad (7)$$

Automated Construction of Knowledge Bases

Automatically constructed Knowledge Bases: Learning new clauses, w neural techniques

$$\mathbf{y}_{e_1} = f(\mathbf{W}\mathbf{x}_{e_1}), \quad \mathbf{y}_{e_2} = f(\mathbf{W}\mathbf{x}_{e_2})$$

f is linear or non-linear, W is a parameter matrix randomly intialized

$$g_r^a(\mathbf{y}_{e_1}, \mathbf{y}_{e_2}) = \mathbf{A}_r^T \begin{pmatrix} \mathbf{y}_{e_1} \\ \mathbf{y}_{e_2} \end{pmatrix} \text{ and } g_r^b(\mathbf{y}_{e_1}, \mathbf{y}_{e_2}) = \mathbf{y}_{e_1}^T \mathbf{B}_r \mathbf{y}_{e_2},$$

Rep. of a binary relation

Models	$\mathbf{B}_r$	$\mathbf{A}_r^T$	Scoring Function
Distance (Bordes et al., 2011)	-	$(\mathbf{Q}_{r1}^T \quad -\mathbf{Q}_{r2}^T)$	$-\|\|g_r^a(\mathbf{y}_{e_1}, \mathbf{y}_{e_2})\|\|_1$
Single Layer (Socher et al., 2013)	-	$(\mathbf{Q}_{r1}^T \quad \mathbf{Q}_{r2}^T)$	$\mathbf{u}_r^T \tanh(g_r^a(\mathbf{y}_{e_1}, \mathbf{y}_{e_2}))$
TransE (Bordes et al., 2013b)	$\mathbf{I}$	$(\mathbf{V}_r^T \quad -\mathbf{V}_r^T)$	$-(2g_r^a(\mathbf{y}_{e_1}, \mathbf{y}_{e_2}) - 2g_r^b(\mathbf{y}_{e_1}, \mathbf{y}_{e_2}) + \|\|\mathbf{V}_r\|\|_2^2)$
NTN (Socher et al., 2013)	$\mathbf{T}_r$	$(\mathbf{Q}_{r1}^T \quad \mathbf{Q}_{r2}^T)$	$\mathbf{u}_r^T \tanh\left(g_r^a(\mathbf{y}_{e_1}, \mathbf{y}_{e_2}) + g_r^b(\mathbf{y}_{e_1}, \mathbf{y}_{e_2})\right)$

$$L(\Omega) = \sum_{(e_1,r,e_2)\in T} \sum_{(e_1',r,e_2')\in T'} \max\{S_{(e_1',r,e_2')} - S_{(e_1,r,e_2)} + 1, 0\}$$

Training objective.

$$T' = \{(e_1', r, e_2) | e_1' \in E, (e_1', r, e_2) \notin T\} \cup \{(e_1, r, e_2') | e_2' \in E, (e_1, r, e_2') \notin T\}$$

T' (negative examples) obtained from T by "corruption"

	FB15k		FB15k-401		WN	
	MRR	HITS@10	MRR	HITS@10	MRR	HITS@10
NTN	0.25	41.4	0.24	40.5	0.53	66.1
Blinear+Linear	0.30	49.0	0.30	49.4	0.87	91.6
TransE (DISTADD)	0.32	53.9	0.32	54.7	0.38	90.9
Bilinear	0.31	51.9	0.32	52.2	**0.89**	92.8
Bilinear-diag (DISTMULT)	**0.35**	**57.7**	**0.36**	**58.5**	0.83	**94.2**

Link-embedding results

Note: Bilinear rep. corresponds to rep reln as a tensor

- Entities are n-vectors, binary predicates are matrices, p(x,y) evaluated as (x^t p y).
- Evaluate short-circuit clauses, via matrix multiply
  $p(X_1,X_{n+1})$:- $p_1(X_1,X_2)$, $p_2(X_2,X_3)$,..., $p_n(X_n,X_{n+1})$
- Accept if the resulting matrix is "close" to the matrix for p.

e.g.

nationality(A,C) :- bornInCity(A,B), cityOfCountry(B,C).

Extracting short-circuit Horn clauses

But this does not work for more complex kinds of clauses, e.g:

dualCitizen(A, C, C1) :- bornIn(A,B), cityOfCountry(B, C1), C != C1, residentOfCountry(C), naturalizedInCountry(C).

Use differentiable logic techniques to extend?

Yang, Yih, He, Gao, Deng, ICLR 2015. *Embedding Entities and Relations for Learning and inference in KBs*

Path-RNN ++: Improved performance for link-completion, clause synthesis

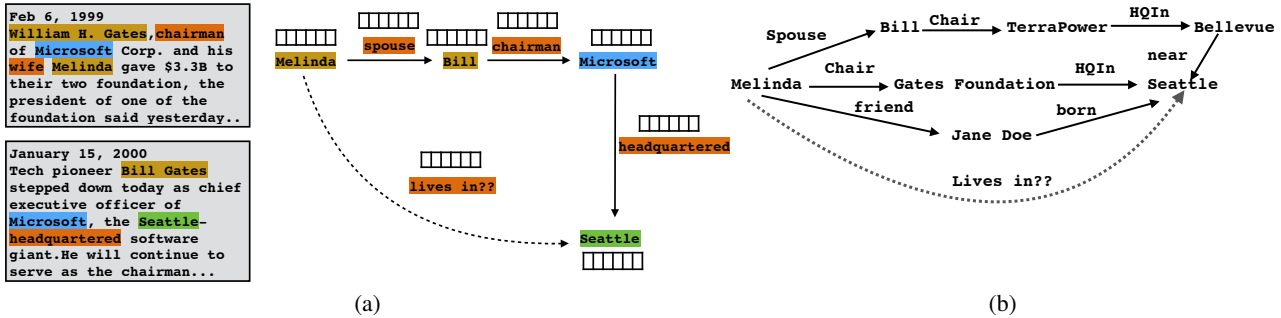
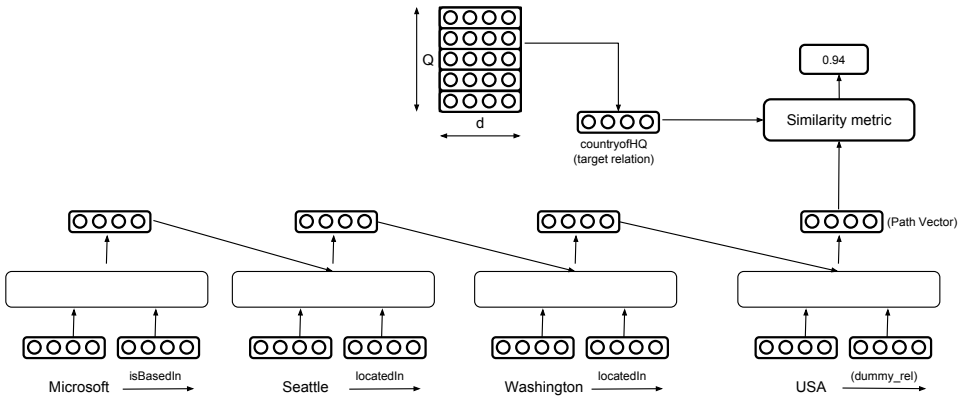


Figure 1: The nodes in the knowledge graphs represent entities and the labeled edges represent relations. (a) A path between 'Melinda' and 'Seattle' combining relations from two different documents. (b) There are multiple paths between entities in a knowledge graph. The top two paths are predictive of the fact that Melinda may '*live in*' Seattle, but the bottom (fictitious) path isn't.

MAP=mean average precision

Model	Performance (%MAP)	Pooling
PRA	64.43	n/a
PRA + Bigram	64.93	n/a
Path-RNN	65.23	Max
Path-RNN	68.43	LogSumExp
Single-Model	**70.11**	LogSumExp
PRA + Types	64.18	n/a
Single-Model	70.11	LogSumExp
Single-Model + Entity	71.74	LogSumExp
Single-Model + Types	**73.26**	LogSumExp
Single-Model + Entity + Types	72.22	LogSumExp



Key features:

- Jointly reason about entities, relations and entity types (more accurate reasoning)
- Learn from multiple paths (logsumexp)
- Use a relation-independent RNN for composition.

Das, Neelkantan, Belanger, McCallum 2017. *Chains of Reasoning over Entities, Relations and Text using RNNs*
Arvind Neelakantan, Roth, McCallum. 2015. *Compositional vector space models for knowledge base completion.*
Ni Lao, Tom Mitchell, and William W. Cohen. 2011. *Random walk inference and learning in a large scale knowledge base.*

Expressing anti-symmetry: Moving to complex embeddings (ComplEx)

$$\langle u, v \rangle := \bar{u}^T v$$

$$\phi(r, s, o; \Theta) \quad = \quad \mathrm{Re}(< w_r, e_s, \bar{e}_o >)$$

Consider normal matrices X: $X\bar{X}^T = \bar{X}^T X$
X is normal iff it is unitarily diagonalizable:

$$X = EW\bar{E}^T$$

Where W is the diagonal matrix of (complex) eigen-values

Basic idea is to move to embeddings over the complex – dot product involves complex conjugate, hence asymmetric.

Scoring function ϕ: Asymmetry handled through complex conjugation of entity embeddings

Regularization by low sign-rank.

Learning through (low-rank) matrix factorization

Dataset	$\lvert\mathcal{E}\rvert$	$\lvert\mathcal{R}\rvert$	#triples in Train/Valid/Test
WN18	40,943	18	141,442 / 5,000 / 5,000
FB15K	14,951	1,345	483,142 / 50,000 / 59,071

	WN18					FB15K				
	MRR		Hits at			MRR		Hits at		
Model	Filter	Raw	1	3	10	Filter	Raw	1	3	10
CP	0.075	0.058	0.049	0.080	0.125	0.326	0.152	0.219	0.376	0.532
TransE	0.454	0.335	0.089	0.823	0.934	0.380	0.221	0.231	0.472	0.641
DistMult	0.822	0.532	0.728	0.914	0.936	0.654	**0.242**	0.546	0.733	0.824
HolE*	0.938	**0.616**	0.93	**0.945**	**0.949**	0.524	0.232	0.402	0.613	0.739
ComplEx	**0.941**	0.587	**0.936**	**0.945**	0.947	**0.692**	**0.242**	**0.599**	**0.759**	**0.840**

Filtered and Raw Mean Reciprocal Rank (MRR)

Trouillon, Welbl, Riedel, Gaussier, Bouchard. Complex Embeddings for Simple Link Prediction. ICML 2016

Neural Theorem Prover for Datalog

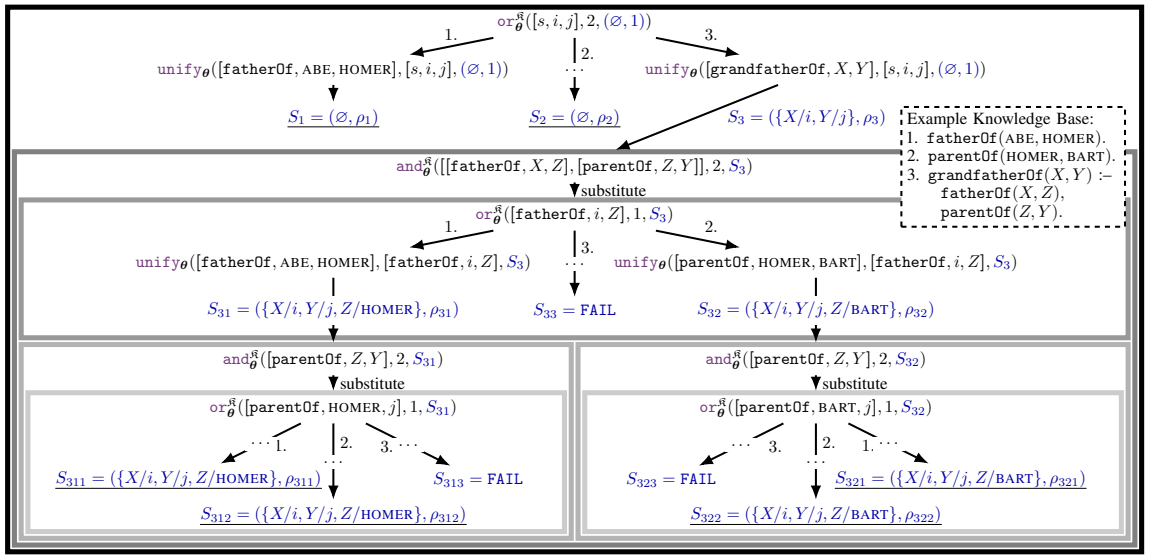


Table 1: AUC-PR results on Countries and MRR and HITS@m on Kinship, Nations, and UMLS.

Corpus		Metric	Model			Examples of induced rules and their confidence
			ComplEx	**NTP**	**NTPλ**	
Countries	S1	AUC-PR	99.37 ± 0.4	90.83 ± 15.4	**100.00** $\pm$ 0.0	0.90 `locatedIn(X,Y) :– locatedIn(X,Z), locatedIn(Z,Y).`
	S2	AUC-PR	87.95 ± 2.8	87.40 ± 11.7	**93.04** $\pm$ 0.4	0.63 `locatedIn(X,Y) :– neighborOf(X,Z), locatedIn(Z,Y).`
	S3	AUC-PR	48.44 ± 6.3	56.68 ± 17.6	**77.26** $\pm$ 17.0	0.32 `locatedIn(X,Y) :– neighborOf(X,Z), neighborOf(Z,W), locatedIn(W,Y).`
Kinship		MRR	0.46	0.36	**0.48**	0.98 `term15(X,Y) :– term5(Y,X)`
		HITS@1	0.34	0.24	**0.39**	0.97 `term18(X,Y) :– term18(Y,X)`
		HITS@3	**0.49**	0.40	0.47	0.86 `term4(X,Y) :– term4(Y,X)`
		HITS@10	**0.74**	0.60	0.71	0.73 `term12(X,Y) :– term10(X, Z), term12(Z, Y).`
Nations		MRR	0.60	**0.63**	0.62	0.68 `blockpositionindex(X,Y) :– blockpositionindex(Y,X).`
		HITS@1	0.46	**0.48**	0.45	0.46 `expeldiplomats(X,Y) :– negativebehavior(X,Y).`
		HITS@3	0.67	0.69	**0.72**	0.38 `negativecomm(X,Y) :– commonbloc0(X,Y).`
		HITS@10	0.97	0.98	**0.99**	0.38 `intergovorgs3(X,Y) :– intergovorgs(Y,X).`
UMLS		MRR	0.58	0.57	**0.60**	0.88 `interacts_with(X,Y) :– interacts_with(X,Z), interacts_with(Z,Y).`
		HITS@1	0.47	0.47	**0.51**	
		HITS@3	0.63	0.60	**0.64**	0.77 `isa(X,Y) :– isa(X,Z), isa(Z,Y).`
		HITS@10	0.80	0.79	**0.81**	0.71 `derivative_of(X,Y) :– derivative_of(X,Z), derivative_of(Z,Y).`

How does one support transitive reasoning in a neural framework?

General approach to “Differentiable Logic Programming”: end-to-end-differentiable theorem prover with sub-symbolic representations

Use dynamic neural network modules for recursive composition: **unify**, **and** and **or** modules. Each module consumes and produces *proof state:* (ψ, ρ) where ψ is a substitution and ρ a neural network that scores partial proofs.

Replace equality check in unification with comparison based on Radial Basis Function kernel.

Induced logic rules are inspectable

Tim Rocktaschel, Sebastian Riedel: End-to-end Differentiable Proving, 2017

Differentiable Logic

Differentiable Logic -- A Precise Embedding of Logic in Linear Algebra

B = {0, 1} I = [0, 1] U = B^n "The Universe"

false true

not x = 1 - x

x and y = x y

Using $X = XX$ all monomials (over $x_1 \ldots x_n$) can be reduced to $\prod_{i \in A} x_i$ (for $A \subseteq \{1, \ldots, n\}$)

➔ Monomials of interest to us are *multilinear*

➔ These are representable as *tensors*

---

x = y = xy + (1-x)(1-y) (xor)

$X = Y = \prod_{i \in 1 \ldots n} X_i = Y_i$

In fact, any function $f: U \rightarrow R$ can be represented by its "Fourier expansion", a formula of size $5n2^n$

$$\lambda X. \sum_{A \in U} f(A)[A = X]$$

Goal is to develop a framework which can support functions with "deep knowledge", that require much less data to be trained. (Leveraging a 100+ years of logic!)

Saraswat, Jagadeesan, 2017. *Notes on Differentiable Logic*

Expressing Predicate and Functional Structure

B = {0, 1} I = [0, 1] U = B^n "The Universe"

false true

A k-ary predicate p or function f can be represented by a formula of size $5kn2^{kn}$:

$$p = \lambda\, X_1, \ldots, X_k\, .\sum_{A_1 \ldots A_k \in U} p(A_1, \ldots, A_k) \prod_{i \in 1 \ldots k} X_i = A_i$$

Universal quantification = Bounded Conjunction
Existential quantification = Bounded Disjunction

Use Indicator Functions!

Saraswat, Jagadeesan, 2017. *Notes on Differentiable Logic*

Differentiable logic: Move to a continuous embedding

B = {0, 1} I = [0, 1] U = ~~B~~^{n} I^n "The Universe"

false true

I^n represents a continuous "smearing" of the space of "perfect" individuals

The *continuous completion* p^c of a formula is its Fourier expansion, with variables ranging over I^n.

- It accurately represents the behavior of p on perfect individuals,
- Linearly interpolates this behavior on smeared points.

A basis for logic where individuals are represented via embeddings, and formulas are differentiable.

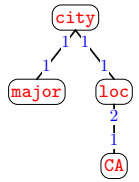
Saraswat, Jagadeesan, 2017. *Notes on Differentiable Logic*

Neural Approaches to DB query using nat lang

Neural Theorem Prover for Datalog

Example: *major city in California*

$z = \langle \texttt{city}; \frac{1}{1}{:}\langle \texttt{major} \rangle ; \frac{1}{1}{:}\langle \texttt{loc}; \frac{2}{1}{:}\langle \texttt{CA} \rangle \rangle \rangle$



$$\lambda c\, \exists m\, \exists \ell\, \exists s\,.\; \texttt{city}(c) \wedge \texttt{major}(m) \wedge \texttt{loc}(\ell) \wedge \texttt{CA}(s) \wedge c_1 = m_1 \wedge c_1 = \ell_1 \wedge \ell_2 = s_1$$

(a) DCS tree (b) Lambda calculus formula

(c) Denotation: $[\![z]\!]_w = \{\texttt{SF}, \texttt{LA}, \ldots\}$

System	GEO	JOBS
Tang and Mooney (2001)	79.4	79.8
Wong and Mooney (2007)	86.6	–
Zettlemoyer and Collins (2005)	79.3	79.3
Zettlemoyer and Collins (2007)	86.1	–
Kwiatkowski et al. (2010)	88.2	–
Kwiatkowski et al. (2010)	88.9	–
Our system (DCS with L)	88.6	91.4
Our system (DCS with L^+)	**91.1**	**95.0**

Key contribution is a semantic form (DCS) that is compact for the queries in this domain.

Learning needs beam search + dynamic programming.

Rules provided (semantic trigger).

This approach is still state of the art on geo-query

Percy Liang, Michael Jordan, Dan Klein “Learning Dependency-based Compositional Semantics”, 2011
Percy Liang PhD thesis 2011

Using RNN-based parsers for semantic parsing

Cheng, Reddy, Saraswat, Lapata “Learning Natural Language Meaning Representations for Semantic Parsing”. ACL 2017

Neural Programmer

Using a NN to synthesize logic form for answering a natural language query against a database (with never-seen-before column headings), using indirect supervision.

Arvind Neelkantan, ICLR 2016, 2017. Neural Programmer.
Arvind Neelkantan PhD Thesis, 2017, U Mass.

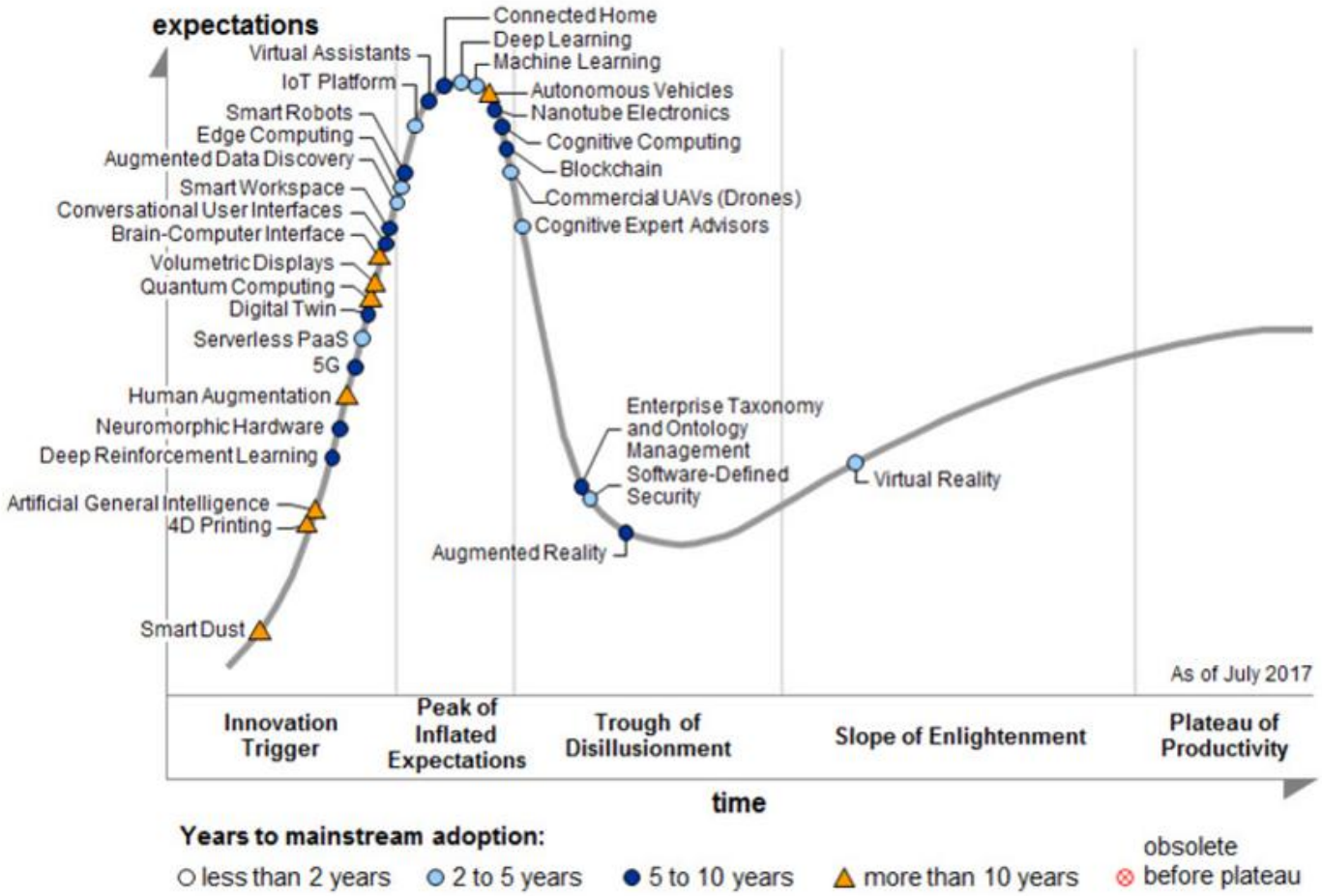
Conclusion

This is a dawn of a new *post-algorithmic* era:

Combining differentiable and symbolic techniques for learning to reason with language and logic, leveraging a large amount of (possibly noisy) data, and throwing (potentially tremendous) compute power at the problem

Deep Learning on the Gartner Hype-cycle: Really?

Hype Cycle for Emerging Technologies, 2017



Note: PaaS = platform as a service; UAVs = unmanned aerial vehicles

Source: Gartner (July 2017)